PSYCHOMETRIKA—VOL. 69, NO. 4, 593–611
DECEMBER 2004

SELECTION OF VARIABLES IN EXPLORATORY FACTOR ANALYSIS: AN EMPIRICAL COMPARISON OF A STEPWISE AND TRADITIONAL APPROACH

KRISTINE Y. HOGARTY
JEFFREY D. KROMREY
JOHN M. FERRON
CONSTANCE V. HINES

UNIVERSITY OF SOUTH FLORIDA

The purpose of this study was to investigate and compare the performance of a stepwise variable selection algorithm to traditional exploratory factor analysis. The Monte Carlo study included six factors in the design; the number of common factors; the number of variables explained by the common factors; the magnitude of factor loadings; the number of variables not explained by the common factors; the type of anomaly evidenced by the poorly explained variables; and sample size. The performance of the methods was evaluated in terms of selection and pattern accuracy, and bias and root mean squared error of the structure coefficients. Results indicate that the stepwise algorithm was generally ineffective at excluding anomalous variables from the factor model. The poor selection accuracy of the stepwise approach suggests that it should be avoided.

Key words: Stepwise variable selection, exploratory factor analysis, goodness-of-fit, varimax rotation, statistical bias, selection accuracy, pattern accuracy.

Selection of Variables in Exploratory Factor Analysis: An Empirical Comparison of Stepwise and Traditional Approaches

Exploratory Factor Analysis (EFA) is among the most widely used statistical methods in psychological research. Despite the popularity of the method, concerns have been raised with respect to the quality of EFAs reported in the psychological literature (Fabrigar, Wegener, MacCallum, & Strahan, 1999). Criticism of the work in this area stem, to a large extent, from questionable methodological decisions made by researchers. Fabrigar et al. (1999) identified five major methodological issues requiring decisions on the part of the researcher in the implementation of exploratory factor analysis. These include (a) the selection of variables to be used in the study, the size and nature of the sample, and consideration of the psychometric properties of the measures, (b) the determination of whether EFA is the most appropriate form of analysis given the goals of the research project, (c) the choice of method to fit the common factor model (e.g., factor extraction procedures), (d) the criteria to employ to determine the number of factors to retain, and (e) the method for rotating factors to yield a final interpretable solution. Failure to make an appropriate decision regarding one or more of these methodological issues may lead to erroneous results and limit the utility of the factor analysis.

Over the years, much attention has been given in the literature to several of the methodological decisions noted above (e.g., methods of factor extraction, the number of factors to retain, and factor rotation methods). In contrast, research methodologists have given limited consideration to variable selection. Little, Lindenberger, and Nesselroade (1999) have suggested that the selection of indicators (or variables) has typically relied on informal or intuitive reasoning or historical precedent. They note that the issue of variable selection is directly related to the quality of the research design and the value of the results.

Requests for reprints should be sent to Kristine Y. Hogarty, Department of Educational Measurement and Research, University of South Florida, 4202 E. Fowler Ave., EDU 162, Tampa, FL 33620. E-mail: hogarty@tempest.coedu.usf.edu

0033-3123/2004-4/2001-0907-A $00.75/0

There are two distinct phases in the conduct of an exploratory factor analysis in which a researcher must make decisions regarding variable selection. The first phase occurs during the development of the research design when decisions are made with respect to which variables from the domain of interest should be included in the study. Of course, the decision regarding what measured variables should be included is of critical importance to a study (Cattell, 1978). Inadequate sampling of measured variables from the domain of interest may lead to failure to uncover important common factors. On the other hand, the inclusion of variables that are irrelevant to the domain of interest may mask the true common factors or lead to the emergence of spurious factors (Fabrigar et al., 1999). Issues central to the selection of variables include both the number of variables to be included in the study and the nature of those variables.

The second phase of variable selection occurs during data analysis, when decisions are made relative to the number of common factors to be retained in the model to yield a parsimonious set of factors that best reflects the underlying constructs being measured and the selection of variables that adequately represent each of these common factors. There is evidence to suggest that the results of exploratory factor analysis are more accurate when each common factor is represented by multiple variables in the analysis. In this regard, MacCallum, Widaman, Zhang, and Hong (1999), and Velicer and Fava (1998) recommend that a minimum of three to five variables represent each common factor. Additional considerations regarding the nature of the variables include both cost and availability, as well as the meaning of the variables and the relationships among them (Tabachnick & Fidell, 1983).

Further, Gorsuch (1988) identified two basic criteria for the selection of variables, namely the reliability of the variables, and the expected correlations with other variables in the analysis. If the reliability of variables is low, or if the correlation of a given variable with the other variables in the domain of interest is low, then the communality of the variable will be low, as it would share little in common with the other variables in the domain. Fabrigar et al. (1999), and MacCallum et al. (1999), posit that when EFA is performed on variables with low communalities substantial distortion in results may occur.

The focus of this paper is on variable selection of the second nature, that is, variable selection that occurs during the data analysis phase of exploratory factor analysis. Because there is a relative lack of formal guidelines for informing critical decisions on variable selection during this phase, there is a compelling need for continuing efforts to explore and develop methods in this regard. In EFA, the magnitude of factor loadings is frequently used as a criterion to determine which variables are substantially related to a given factor and thus should be retained. Several "rules of thumb" are presented in the literature to provide guidance in this area. For example, while the .30 loading magnitude is one of the most popular criteria for the interpretation of factor analysis results (e.g., Nunnally, 1994; Tabachnick & Fidell, 1983), Comrey (1973) offers the additional suggestion that loadings in excess of .71 are excellent, .63 are very good, .55 are good, .45 fair, and .32 poor. These rules, however, do not represent an undisputed choice, and often researchers choose different values based on other preferences. According to Cudeck and O'Dell (1994), although a variety of rules of thumb of this nature are venerable, they are often ad hoc and ill advised. For example, the practice of scanning coefficients for the largest loading while disregarding the others is sometimes employed, but such a criterion ignores the possibility of any complexity in the neglected variables. An alternative method noted by these authors is to interpret the loadings in the context of their standard errors (using a rough approximation for the standard error such as the reciprocal of the square root of N). Clearly, simple rules of thumb do not make allowances for the influence of many other important factors, such as estimation method, factor rotation, number of factors, sample size, and the clarity of the factor solution (Cudeck & O'Dell, 1994).

In an attempt to find a more objective method for making decisions about variable selection in factor analysis, Kano and Harada (2000) proposed the use of goodness-of-fit statistics within the context of stepwise variable selection procedures. Because of the controversy surrounding the use of stepwise procedures and the paucity of research regarding the use of these methods for factor analysis, the primary focus of this study was to examine the behavior of this proposed stepwise algorithm developed for variable selection (Kano & Harada, 2000).

The use of stepwise methods for the selection of a subset of variables or to evaluate the order of importance of variables has been the subject of much criticism in the literature (see for example Huberty, 1989; Snyder, 1991; Thompson, 1988b, 1989). Three commonly stated criticisms surrounding conventional stepwise methods include a dramatic inflation of Type I error rates, the failure to identify the best set of predictor variables of a particular size, and a tendency to capitalize on sampling error, yielding conclusions that often are not replicable. Although previous indictments of stepwise approaches have focused primarily on regression analysis (Thompson, 1988b, 1989, 1995) and discriminant analysis (Huberty, 1989), concerns with data driven approaches to model modification can be found elsewhere. For example, in structural equation modeling similar concerns have been raised about the use of modification indices (MacCallum, Roznowski, & Necowitz, 1992), which allow researchers to free and constrain parameters in a model. Since many analytic methods are correlational and are related (Knapp, 1978; Thompson, 1988a) these concerns may generalize across the family of commonly applied correlational methods.

In this study the performance of a stepwise selection algorithm was investigated and compared to a traditional approach to variable selection in exploratory factor analysis using Monte Carlo methods. The majority of variables included in the design phase of the study were consistent with a pre-determined common factor model. However, anomalous variables were also considered to facilitate assessment of the extent to which each of the approaches to variable selection was effective in excluding anomalous variables from the resultant common factor model.

A Stepwise Variable Selection Algorithm

Consider a p-variate sample covariance matrix, **S**, that is hypothesized to result from an underlying common factor model (Thurstone, 1947). That is, the observed values of the p-variables, **x**, can be modeled as

$$\mathbf{x} = \boldsymbol{\mu} + \Lambda\mathbf{f} + \mathbf{u} \quad (1)$$

where

$\boldsymbol{\mu}$ = a vector of population means

Λ = a $p \times k$ matrix of factor loadings

$\mathbf{f}$ = a vector of scores on k common factors, and

$\mathbf{u}$ = a vector of scores on p unique factors.

When the common factor model underlies the data, and the Var $(\mathbf{f}) = \mathbf{I}_k$, Cov $(\mathbf{f}, \mathbf{u}) = \mathbf{0}$, and the Var $(\mathbf{u}) = \Psi$, then the covariance matrix of **x** will be

$$\text{Var}(\mathbf{x}) = \Lambda\Lambda' + \Psi \quad (2)$$

The parameters in Λ and Ψ can be estimated, and these estimates can be used to obtain an implied covariance matrix for **x**,

$$\hat{\Sigma} = \hat{\Lambda}\hat{\Lambda}' + \hat{\Psi} \quad (3)$$

The implied covariance matrix can then be compared to the observed covariance matrix, **S**, to test the adequacy of a common factor model. More specifically, a goodness of fit χ^2 statistic based on the likelihood ratio test (LRT) can be computed

$$\mathrm{LRT} = n(\log|\hat{\Sigma}| - \log|S|) \tag{4}$$

If the covariance matrix implied by the model is statistically significantly different from the observed covariance matrix, the k factor model can be considered inadequate.

If the k factor model is found to be inadequate, the researchers may wonder if the inadequacy of the model is arising from an anomalous variable. The researchers could remove the first variable, and then check for congruence between the covariance matrix of the remaining variables, $\mathbf{S}_{p-1}$, and the covariance matrix implied by the k factor model, $\hat{\Sigma}_{p-1}$. This process could be repeated for each of the p variables. The researchers could then remove the variable that would most improve the fit between the k factor model and the data. If the test statistic was still statistically significant indicating the model was inadequate, the researchers could then remove another variable. The process of removing variables could continue until a subset of variables was found that was consistent with the k factor model.

Conversely, researchers working with a set of p variables for which the k factor model fits adequately, may wish to entertain the possibility of including additional variables. The researchers could add a single variable and then check for congruence between the sample covariance matrix, $\mathbf{S}_{p+1}$, and the covariance matrix implied by the k factor model, $\hat{\Sigma}_{p+1}$. The researchers could use this method to check each of the variables that they are considering adding. The variable leading to the smallest test statistic when $\mathbf{S}_{p+1}$ and $\hat{\Sigma}_{p+1}$ are compared would be added, as long as the resulting test statistic was still not statistically significant, which would indicate the k factor model adequately fit the data. This process could then be repeated to include additional variables.

Kano and Harada (2000) developed a stepwise algorithm that integrates these steps for eliminating and adding variables. To ease the computational demands, a Lagrange Multiplier Test (LMT) was developed and substituted for the LRT. Consideration of this algorithm reveals both similarities and differences with other stepwise procedures. Like other stepwise procedures, one would anticipate that this algorithm would capitalize on sampling error as it steps through the data using statistical tests that do not adequately account for the degree to which the data are being examined. Also like other stepwise procedures, each step is conditional on the results of previous steps. This leaves one wondering if a better subset of variables could have been found by considering all possible subsets. Although these similarities would lead one to be skeptical of the performance of the stepwise factor analysis, there is a difference between stepwise factor analysis and other stepwise algorithms that should also be considered.

The stepwise factor analysis seeks to maximize the number of variables in the model under the constraint that the covariance matrix of the variables does not differ significantly from the covariance matrix implied by a k factor model. The notion is that more indicators are better as long as they are consistent with the model. Since exclusion is based on obtaining statistical significance, one would anticipate that low power would lead to the inclusion of anomalous variables. In contrast, both stepwise regression and stepwise discriminant function analysis are designed such that inclusion, not exclusion, is based on obtaining statistical significance. Thus with these other stepwise procedures one would anticipate that low power would result in the failure to include key variables.

Also note that it can be anticipated that the stepwise algorithm would only be able to identify some kinds of anomalous variables. In particular, the algorithm looks for variables that are inconsistent with a k factor model. If we have p variables that are consistent with a k factor model, then $\hat{\Sigma}_p - \mathbf{S}_p$ approaches $\mathbf{0}$ as n approaches ∞. If we then add a variable such that $\hat{\Sigma}_{p+1} - \mathbf{S}_{p+1}$

does not approach $\mathbf{0}$ as n approaches ∞, we have added a variable that is inconsistent with the k factor model, and would suspect that the stepwise algorithm would identify the variable given an ample sample size. There are variables, however, that are consistent with a k factor model, $\hat{\Sigma}_{p+1} - \mathbf{S}_{p+1}$ approaches $\mathbf{0}$ as n approaches ∞, but that are also a poor choice for a factor analytic study. It would appear that the stepwise algorithm would have difficulty identifying this type of anomalous variable.

As an example, consider a variable i that is uncorrelated with p variables that are consistent with a k factor model. The population covariance matrix,

$$\Sigma_{p+1} = \begin{bmatrix} \Sigma_p & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\sigma}_i^2 \end{bmatrix},$$

can be obtained from model parameters,

$$\Sigma_{p+1} = \Lambda_{p+1}\Lambda'_{p+1} + \Psi_{p+1},$$

where

$$\Lambda_{p+1} = \begin{bmatrix} \Lambda_p \\ \mathbf{0} \end{bmatrix} \quad \text{and} \quad \Psi_{p+1} = \begin{bmatrix} \Psi_p & \\ & \boldsymbol{\sigma}_i^2 \end{bmatrix}.$$

For this type of anomalous variable $\hat{\Sigma}_{p+1} - \mathbf{S}_{p+1}$ approaches $\mathbf{0}$ as n approaches ∞, making the variable difficult to identify. As a second example consider a variable i that has a relatively small level of relationship, say pattern coefficients less than .3, with each of the k factors. Again the population covariance matrix could be obtained from the model parameters,

$$\Sigma_{p+1} = \Lambda_{p+1}\Lambda'_{p+1} + \Psi_{p+1},$$

where

$$\Lambda_{p+1} = \begin{bmatrix} \Lambda_p \\ \mathbf{a} \end{bmatrix}, \quad \Psi_{p+1} = \begin{bmatrix} \Psi_p & \\ & \boldsymbol{\psi}_i^2 \end{bmatrix},$$

$\mathbf{a}$ is a vector of small pattern coefficients, and $\boldsymbol{\psi}_i^2$ is the unique variance of variable i. One would anticipate that this type of anomalous variable would also be difficult to detect using the stepwise algorithm.

Method

The performance of the stepwise algorithm was investigated using Monte Carlo methods, in which random samples were generated under known and controlled population conditions. In the Monte Carlo study, samples were generated from known populations and exploratory factor analyses were conducted on each sample. The conditions for the study were designed to model the situation in which a researcher obtains observations on a sample of $p + q$ variables with the intent to estimate k common factors (i.e., the number of factors is determined *a priori*). The $p + q$ variables represent a “menu” of possible variables to be included in estimating the final factor solution. However, only p variables are well explained by the k common factors. The remaining q variables have population correlations with the p variables that are low, zero, or not well explained with the k common factors.

The Monte Carlo study included six factors in the design. These factors were (a) the number of common factors, k, present in the population (populations were simulated with $k = 2$, and 5 uncorrelated factors), (b) the number of variables, p, that are well explained by the common factors (with $p = 3k$, $5k$, and $10k$), (c) the factor loadings of the p well explained variables (with

$\lambda = .4, .6$, and $.8$), (d) the number of variables, q, that are not well explained by the common factors (with $q = 1$ and k), (e) the type of anomaly evidenced by the q poorly explained variables (no correlation with the other variables, low level of correlation dispersed among all the other variables, and focused correlation inconsistent with the k factor model), and (f) sample size (with samples of $5(p+q)$, $10(p+q)$, and $50(p+q)$). Crossing the six factors leads to 324 conditions. The combination of p and q variables provide population conditions that range from those with 2% of the items being anomalous (with $p = 10k$, $k = 5$, and $q = 1$) to those with 25% of the items being anomalous (with $p = 3k$, $k = 2$, and $q = k$).

Traditional approaches and recent practices using factor analytic methods were used to guide the selection of the aforementioned factors and conditions. For example, Gorsuch (1973) suggested a ratio of 5 participants per measured variable, but a sample size never less than 100, while Nunnally (1994) and Everitt (1975) proposed ratios of 10 to 1. These guidelines were considered and coupled with the findings of more recent research (MacCallum et al., 1999; Velicer & Fava, 1998) suggesting the inclusion of other important characteristics of the data such as the extent to which factors are overdetermined and the level of communality.

Because the extant literature provides no empirical evidence regarding the performance of the stepwise method, the conditions examined in the Monte Carlo study were selected to represent the simplest data structures for exploratory factor analysis. For this reason, the samples were generated from populations characterized by orthogonal factors, completely simple structure, and a consistent number of variables loading on each factor.

The samples were generated by first constructing a population covariance matrix for each condition simulated, then generating random samples from that population. The population correlation matrix was defined as

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{pp} & \mathbf{R}_{pq} \\ \mathbf{R}_{qp} & \mathbf{R}_{qq} \end{bmatrix}, \tag{5}$$

where

$$\begin{aligned} \mathbf{R}_{pp} &= \Lambda\Lambda' + \omega, \\ \Lambda &= \text{the } p \times k \text{ population factor structure matrix}, \\ \Psi &= \text{the } p \times p \text{ diagonal matrix of unique variances}, \\ \mathbf{R}_{qq} &= \text{a } q \times q \text{ symmetric correlation matrix, and} \\ \mathbf{R}_{qp} &= \text{a } p \times q \text{ asymmetric matrix of correlations.} \end{aligned}$$

The elements of Λ were set to $\lambda(= .4, .6$, or $.8)$ for one of the factors and were set to zero for the other factors (i.e., representing a perfectly simple factor structure in the population for the p variables). The elements of $\mathbf{R}_{qq}$ and $\mathbf{R}_{qp}$ depended on the type of anomalous variables. For conditions containing variables with no commonality $\mathbf{R}_{qq} = \mathbf{I}$ and $\mathbf{R}_{qp} = \mathbf{0}$. For conditions with low-level dispersed correlations, the elements of $\mathbf{R}_{qq}$ and $\mathbf{R}_{qp}$ were generated from $\mathbf{R} = \Lambda\Lambda' + \Psi$, setting the elements of Λ corresponding to the q variables to

$$\left(\frac{h^2}{k}\right)^{1/2},$$

with $h^2 = .02$. Finally, for conditions with focused correlations that could not be explained by the k common factors, $\mathbf{R}_{qq} = \mathbf{I}$ and each row of $\mathbf{R}_{qp}$ contained a single nonzero element of .2, with the constraint that no two rows could have the same column with a nonzero element.

The research was conducted using SAS/IML version 8. Conditions for the study were run under Windows 98. Normally distributed random variables were generated using the RANNOR random number generator in SAS. A different seed value for the random number generator was used in each execution of the program. The program code was verified by hand-checking results from benchmark datasets. For each condition investigated, 10,000 samples were generated. The use of 10,000 samples provides adequate precision for the investigation of the sampling behavior of the variable selection algorithm. For example, 10,000 samples provide a maximum 95% confidence interval width around an observed proportion that is $\pm.0098$ (Robey & Barcikowski, 1992).

In each sample, the stepwise algorithm proposed by Kano and Harada (2000) was run. The algorithm was started using principal axis factor extraction based on the p variables that are well explained by the k common factors. Squared multiple correlation coefficients (SMCs) were used as initial communality estimates. Following the stepwise algorithm's use of LMTs for variable inclusion and exclusion decisions, the procedure concluded with a varimax rotation of the solution based on the selected variables. As a comparison procedure, a traditional least-squares exploratory factor analysis was conducted on each sample, using all $p + q$ variables without incorporating any variable selection procedure. The traditional procedure used principal axis factor extraction with SMCs as initial communality estimates and a varimax rotation of the k factor solution.

The relative performance of the stepwise algorithm was evaluated by comparing the obtained loading matrix to the population loading matrix. Since the order of the factors (columns) in the obtained loading matrix may differ from the ordering in the population matrix, an algorithm was written to reorder the columns in the obtained loading matrix to minimize the squared differences with the population loading matrix. Several indices were then developed to summarize the differences in the matrices.

1. Sensitivity was calculated as the proportion of samples that yielded the correct selection of the p variables that are well explained by the factors.
2. Specificity was calculated as the proportion of samples that yielded the correct exclusion of the q variables that are not well explained by the factors.
3. Selection accuracy was defined as the proportion of samples that yielded the correct inclusion of the p variables and the correct exclusion of the q variables.
4. Pattern accuracy was determined by the proportion of samples in which the p variables loaded on the correct factor.

According to Tabachnick and Fidell (1983), as a general rule of thumb, loadings in excess of .30 are eligible for interpretation, whereas lower ones are not. A factor loading of .30 is indicative of a 9% overlap in variance between the variable and the factor. The choice of a "rule of thumb" regarding the size of a loading to be interpreted is, of course, a matter of researcher preference. When employing our comparative procedure, an item was considered to be "selected" if it had at least one factor loading $\geq .30$ in absolute value. For both procedures, an included variable was considered to have loaded on the correct factor if the absolute value of its loading on that factor was $\geq .30$, regardless of magnitudes of loadings on other factors.

In addition, the bias and RMSE of the sample factor structure coefficients resulting from the stepwise algorithm were estimated. For comparative purposes, bias and RMSE were also calculated based on the loading matrix that resulted from the traditional analysis using all the variables with no variable selection procedure. The estimate of statistical bias for the coefficient of the jth variable with the kth factor, $\hat{\lambda}_{jk}$, is given by

$$\text{Bias}\ (\hat{\lambda}_{jk}) = \frac{1}{M}\sum_{m=1}^{M}(\hat{\lambda}_{jkm} - \lambda_{jk}), \tag{6}$$

where $\hat{\lambda}_{jkm}$ = the coefficient obtained from the mth sample, λ_{jk} = the population parameter, and the summation is over the M samples included in the Monte Carlo study.

The estimate of RMSE is given by

$$\text{RMSE}\,(\hat{\lambda}_{jk}) = \left(\frac{1}{M}\sum_{m=1}^{M}(\hat{\lambda}_{jkm} - \lambda_{jk})^2\right)^{1/2} \tag{7}$$

where the elements are as defined above.

Of the indices examined, selection accuracy is most consistent with the purpose behind the development of the stepwise algorithm and thus is central to its evaluation. The indices of sensitivity and specificity support the examination of selection accuracy, and are useful in isolating the source of difficulty in circumstances where the selection accuracy is not adequate. In situations where selection accuracy is adequate, attention can be turned to pattern accuracy, which is central to the purposes motivating a factor analysis. The estimates of bias and RMSE are computed to help provide insight if pattern accuracy is not high, and to provide a basis for choosing between the stepwise and more traditional method if pattern accuracy is comparable and acceptable for both.

Results

To save space, detailed results are provided only for conditions with moderate population factor loadings ($\lambda = .6$). The pattern of results was consistent across the other λ values examined, although more favorable results were obtained in general with $\lambda = .8$ and less favorable results with $\lambda = .4$. The complete results may be obtained from the authors.

Selection Accuracy

The estimates of selection accuracy and its two components (sensitivity and specificity) obtained for the stepwise (SW) and traditional (TA) methods are presented in Tables 1–3. For experimental conditions with q variables that evidence no correlation with the p variables (Table 1), the SW method provided sensitivity values of .95 or higher across all conditions. In contrast, the sensitivity of the TA method varied from .79 to .96, with sensitivity increasing as a function of both the number of variables and the sample size. For example, with $k = 2$, $p/k = 3$, and $q = 2$, the sensitivity of TA ranged from .79 (with $n = 5$ per variable) to .94 (with $n = 50$ per variable). With larger matrices, the sensitivity was higher across sample sizes. With $k = 5$, $p/k = 10$, and $q = 5$, for example, the sensitivity ranged from .94 to .96 across the three sample sizes examined.

In contrast to these sensitivity results, the SW method evidenced a lack of specificity across the experimental conditions (ranging only from .00 to .04). The TA method, however, showed a specificity pattern that was consistent with the sensitivity results, increasing with both sample size and the number of variables. For example, with $k = 2$, $p/k = 5$, and $q = 2$, the specificity of TA ranged from .69 (with $n = 5$ per variable) to .95 (with $n = 50$ per variable). With $k = 5$, $p/k = 10$, and $q = 5$, the sensitivity was high across sample sizes, ranging from .94 to .96.

With specificity and sensitivity combined into selection accuracy, the SW method performed very poorly, with selection accuracy ranging from only .00 to .03, while the TA method's selection accuracy ranged from .27 to .96. As with the sensitivity and specificity results, the selection accuracy of the TA method increased dramatically with larger sample sizes and with larger matrices.

The selection accuracy results for conditions in which the q variables evidenced low, diffuse correlations with the p variables are presented in Table 2. The overall pattern of results with this type of anomalous variable was consistent with that obtained with the uncorrelated q variables.

TABLE 1.
Selection accuracy for stepwise and traditional factor analytic methods [anomaly = no correlation].

				Sensitivity		Specificity		Selection accuracy	
k	*P*/*K*	*Q*	*N*/Var	SW	TA	SW	TA	SW	TA
2	3	1	5	98	79	3	48	3	42
2	3	1	10	98	90	3	79	3	75
2	3	1	50	97	94	3	94	3	94
2	3	2	5	99	79	0	30	0	27
2	3	2	10	98	91	0	71	0	69
2	3	2	50	98	94	0	94	0	94
2	5	1	5	96	87	3	78	3	72
2	5	1	10	97	94	3	93	2	92
2	5	1	50	97	95	2	95	2	95
2	5	2	5	97	89	1	69	0	65
2	5	2	10	97	95	0	92	0	92
2	5	2	50	97	95	0	95	0	95
2	10	1	5	95	93	3	92	3	92
2	10	1	10	96	95	3	95	2	95
2	10	1	50	97	96	2	96	2	96
2	10	2	5	96	93	1	91	1	91
2	10	2	10	97	95	1	95	1	95
2	10	2	50	97	95	1	95	1	95
5	3	1	5	98	89	2	57	2	53
5	3	1	10	97	95	2	92	2	91
5	3	1	50	95	93	3	93	2	93
5	3	5	5	99	88	0	17	0	17
5	3	5	10	99	95	0	86	0	86
5	3	5	50	98	93	0	93	0	93
5	5	1	5	96	94	3	91	2	90
5	5	1	10	97	95	2	95	2	95
5	5	1	50	97	96	2	96	2	96
5	5	5	5	98	95	1	84	1	84
5	5	5	10	98	95	1	95	1	95
5	5	5	50	98	95	1	95	1	95
5	10	1	5	96	94	4	94	3	94
5	10	1	10	97	96	3	96	3	96
5	10	1	50	97	96	3	96	2	96
5	10	5	5	97	94	2	94	2	94
5	10	5	10	98	96	2	96	2	96
5	10	5	50	98	96	1	96	1	96

Note. Values have been multiplied by 100 and rounded. Estimates are based on 10,000 replications.

TABLE 2.
Selection accuracy for stepwise and traditional factor analytic methods [anomaly = low level dispersed correlation].

				Sensitivity		Specificity		Selection accuracy	
k	P/K	Q	N/Var	SW	TA	SW	TA	SW	TA
2	3	1	5	98	79	3	43	3	36
2	3	1	10	98	91	3	69	3	66
2	3	1	50	97	94	3	94	3	94
2	3	2	5	98	80	0	23	0	21
2	3	2	10	98	91	0	55	0	53
2	3	2	50	98	94	0	94	0	94
2	5	1	5	96	87	3	68	3	62
2	5	1	10	96	95	2	86	2	86
2	5	1	50	97	95	2	95	2	95
2	5	2	5	97	89	0	53	0	49
2	5	2	10	97	95	0	80	0	80
2	5	2	50	97	95	0	95	0	95
2	10	1	5	95	94	3	88	3	87
2	10	1	10	96	94	2	94	2	94
2	10	1	50	97	96	2	96	2	96
2	10	2	5	96	94	1	83	1	82
2	10	2	10	97	95	1	94	1	94
2	10	2	50	97	95	1	95	1	95
5	3	1	5	98	89	2	50	2	46
5	3	1	10	97	95	2	85	2	85
5	3	1	50	96	94	3	94	2	94
5	3	5	5	99	89	0	10	0	10
5	3	5	10	98	95	0	72	0	71
5	3	5	50	98	94	0	94	0	94
5	5	1	5	96	95	2	86	2	86
5	5	1	10	96	95	2	95	2	95
5	5	1	50	97	96	2	96	2	96
5	5	5	5	98	95	1	69	1	69
5	5	5	10	98	95	1	94	1	94
5	5	5	50	98	95	1	95	1	95
5	10	1	5	96	94	4	94	3	94
5	10	1	10	96	96	3	95	2	95
5	10	1	50	97	96	2	96	2	96
5	10	5	5	97	94	2	94	2	94
5	10	5	10	97	95	2	95	2	95
5	10	5	50	98	96	1	96	1	96

Note. Values have been multiplied by 100 and rounded. Estimates are based on 10,000 replications.

That is, the SW method evidenced extremely high sensitivity but near absence of specificity across the conditions examined. Also consistent with the previous results, the TA method showed increasing specificity and sensitivity with larger matrices, larger samples, and larger population factor loadings.

Finally, the selection accuracy results for conditions in which the q variables evidenced focused but inconsistent correlations with the p variables are presented in Table 3. With this type of anomalous variable, the SW method showed improved performance. The sensitivity of SW remained high across the conditions examined, but the method showed an increase in specificity as sample size increased. For example, with $k = 2$, $p/k = 5$, and $q = 2$, the specificity of SW increased from .02 (with $n = 5$ per variable) to .78 (with $n = 50$ per variable). The SW method also evidenced increased specificity with larger numbers of variables. With $k = 5$, $p/k = 10$, and $q = 5$, for example, the specificity ranged from .03 (with $n = 5$ per variable) to .87 (with $n = 50$ per variable), somewhat greater specificity than was evident with smaller matrices. The performance of the TA method with this type of anomalous variable was consistent with that obtained with the other anomalies examined. That is, both the sensitivity and the specificity of TA increased with larger sample sizes and larger numbers of variables. The overall selection accuracy results (combining sensitivity and specificity) ranged from .01 to .88 for SW and from .05 to .96 for TA. The largest differences between methods were evident with moderate sample sizes ($n = 10$ per variable) and a trend to convergence was seen with large samples.

Pattern Accuracy

The estimates of pattern accuracy obtained for the SW and TA methods are presented in Tables 4–6. These estimates reflect the percentage of samples in which the p variables, those variables well explained by the common factors, loaded on the correct factor. Two different estimates of pattern accuracy are presented. The first estimate of pattern accuracy (per item), represents an average of the proportion of times each of the p variables loaded correctly on the k factors, the second estimate (all items), represents the proportion of samples in which all of the p variables loaded correctly on the k factors. Both methods evidenced a relatively high and consistent degree of pattern accuracy on a per item basis. The per item pattern accuracy exceeded .90 for all conditions examined, with estimates reaching 1.0 with larger matrices, larger sample sizes, and larger number of factors.

Both the SW and TA evidenced consistent and increasing pattern accuracy for all items, across the various anomalous conditions, with larger sample sizes and larger matrices. For most of the conditions examined, the SW method evidenced higher overall pattern accuracy, but the difference in the two methods was relatively minor. As might be expected, the overall pattern accuracy was poorest with smaller sample sizes, smaller numbers of factors, and a lower p/k ratio. For example, for experimental conditions with q variables that evidence no correlation with the p variables (Table 4), with $k = 2$ and $p/k = 3$, the pattern accuracy ranged from .69 to .98 for SW and from .68 to .94 for TA. For the same condition, but with $p/k = 10$, the accuracy ranged from .95 to .97 for SW and from .93 to .96 for TA.

Statistical Bias

The estimates of statistical bias in the factor structure coefficients obtained for the SW and TA methods are also presented in Tables 4–6. These estimates, the average bias in sample structure coefficients, were calculated for both the p variables and the q variables separately. In addition, an overall estimate of statistical bias for each of the conditions under study was computed. For all conditions examined, the estimated bias in the coefficients of the p variables was negligible for both SW and TA methods (with no bias value exceeding .02); however, differences were evident in the bias of the structure coefficients for the q variables.

TABLE 3.
Selection accuracy for stepwise and traditional factor analytic methods [anomaly = focused inconsistent correlation].

				Sensitivity		Specificity		Selection accuracy	
k	P/K	Q	N/Var	SW	TA	SW	TA	SW	TA
2	3	1	5	97	77	7	39	7	34
2	3	1	10	96	89	13	66	12	63
2	3	1	50	94	94	76	93	73	93
2	3	2	5	98	77	1	20	1	18
2	3	2	10	97	89	3	52	3	50
2	3	2	50	94	94	69	93	68	93
2	5	1	5	95	87	7	73	6	67
2	5	1	10	96	95	14	90	13	90
2	5	1	50	95	95	84	95	80	95
2	5	2	5	96	88	2	59	2	56
2	5	2	10	96	95	5	88	5	88
2	5	2	50	95	95	78	95	74	95
2	10	1	5	94	93	9	92	7	91
2	10	1	10	95	95	15	95	12	95
2	10	1	50	96	96	89	96	85	96
2	10	2	5	94	93	4	91	3	90
2	10	2	10	95	95	8	95	7	95
2	10	2	50	95	95	85	95	80	95
5	3	1	5	97	88	4	42	4	39
5	3	1	10	96	95	11	81	10	81
5	3	1	50	94	94	86	94	81	94
5	3	5	5	97	84	0	6	0	5
5	3	5	10	96	95	1	58	1	57
5	3	5	50	93	93	79	93	75	93
5	5	1	5	95	94	7	87	6	86
5	5	1	10	95	95	14	95	12	95
5	5	1	50	95	95	89	95	85	95
5	5	5	5	94	94	1	73	1	73
5	5	5	10	95	95	3	94	3	94
5	5	5	50	95	95	83	95	79	95
5	10	1	5	95	94	10	94	8	94
5	10	1	10	95	95	16	95	13	95
5	10	1	50	96	96	92	96	88	96
5	10	5	5	94	94	3	94	3	94
5	10	5	10	96	96	8	96	6	96
5	10	5	50	96	96	87	96	83	96

Note. Values have been multiplied by 100 and rounded. Estimates are based on 10,000 replications.

TABLE 4.
Statistical bias, RMSE and pattern accuracy for stepwise and traditional exploratory factor analyses [anomaly = no correlation].

				Bias						RMSE						Pattern accuracy			
				p Vars		q Vars		Overall		p Vars		q Vars		Overall		Per item		All items	
k	p/k	q	n/Var	SW	TA	SW	TA	SW	TA	SW	TA	SW	TA	SW	TA	SW	TA	SW	TA
2	3	1	5	−2	−1	−1	0	−2	−1	18	18	20	22	19	20	93	92	69	68
2	3	1	10	−2	−2	0	8	−2	0	11	12	17	32	12	15	98	98	91	89
2	3	1	50	−2	−2	−2	5	−2	2	6	7	6	15	6	11	99	98	97	94
2	3	2	5	−2	−1	−1	0	−1	0	17	17	20	23	19	21	93	93	72	69
2	3	2	10	−2	−2	0	4	−2	0	11	11	16	30	12	16	99	98	93	90
2	3	2	50	−2	−2	−1	10	−2	5	6	7	6	20	6	15	100	98	98	94
2	5	1	5	0	0	0	1	0	1	12	12	13	16	12	14	99	98	87	85
2	5	1	10	−1	−1	0	23	−1	2	8	9	12	40	8	11	100	99	96	94
2	5	1	50	−1	−1	−1	4	−1	2	4	4	4	11	4	8	100	99	97	95
2	5	2	5	0	0	0	1	0	1	11	12	13	17	12	15	99	98	90	87
2	5	2	10	−1	−1	0	13	0	2	8	8	11	33	8	12	100	99	97	94
2	5	2	50	−1	−1	−1	8	−1	4	4	4	4	15	4	11	100	99	97	95
2	10	1	5	0	0	0	2	0	1	8	8	8	11	8	10	100	99	95	93
2	10	1	10	0	0	0	30	0	1	6	6	8	43	6	8	100	100	96	95
2	10	1	50	0	0	0	3	0	1	3	3	3	6	3	5	100	100	97	96
2	10	2	5	0	0	0	3	0	1	8	8	8	12	8	10	100	99	95	93
2	10	2	10	0	0	0	26	0	2	5	6	8	42	6	9	100	100	97	95
2	10	2	50	0	0	0	5	0	3	2	3	3	9	3	6	100	100	97	95
5	3	1	5	−1	0	0	1	0	0	11	11	17	22	12	12	98	97	81	79
5	3	1	10	−1	−1	0	6	−1	0	7	8	11	23	8	9	100	99	97	95
5	3	1	50	−1	−1	0	11	−1	0	4	4	5	25	4	6	100	99	95	93
5	3	5	5	−1	−1	0	1	0	0	11	11	16	20	12	13	97	97	79	77
5	3	5	10	−1	−1	0	2	−1	0	7	7	10	18	8	10	100	99	98	95
5	3	5	50	−1	−1	0	11	−1	1	4	4	4	24	4	8	100	99	98	93
5	5	1	5	0	0	0	7	0	0	8	8	11	25	8	9	100	99	95	94
5	5	1	10	0	0	0	12	0	0	6	6	8	27	6	7	100	99	97	95
5	5	1	50	0	0	0	12	0	0	3	3	3	26	3	4	100	99	97	96
5	5	5	5	0	0	0	2	0	0	8	8	11	19	8	10	100	99	98	95
5	5	5	10	0	0	0	7	0	1	5	6	7	20	6	8	100	99	98	95
5	5	5	50	0	0	0	12	0	1	3	3	3	22	3	5	100	99	98	95
5	10	1	5	0	0	0	12	0	0	5	6	7	28	5	6	100	100	96	94
5	10	1	10	0	0	0	12	0	0	4	4	5	27	4	4	100	100	97	96
5	10	1	50	0	0	0	12	0	0	2	2	2	27	2	2	100	100	97	96
5	10	5	5	0	0	0	5	0	0	5	5	7	21	5	7	100	100	97	94
5	10	5	10	0	0	0	12	0	0	4	4	5	26	4	5	100	100	98	96
5	10	5	50	0	0	0	12	0	0	2	2	2	24	2	3	100	100	98	96

Note. Values have been multiplied by 100 and rounded. Estimates are based on 10,000 replications.

For experimental conditions with q variables that evidence no correlation with the p variables (Table 4), the TA method evidenced greater statistical bias than the SW method, with bias ranging from .00 to .30. The statistical bias in the TA method was largest with $k = 2$, $p/k = 10$, and moderate sample sizes ($n = 10$ per variable). The bias estimates for conditions in which the q variables evidenced low, diffuse correlations with the p variables are presented in Table 5. The overall pattern of results with this type of anomalous variable was consistent with that obtained with the uncorrelated q variables. That is, the SW method provided unbiased estimates across nearly all of the conditions examined. Also consistent with the previous results, the TA method showed relatively greater bias with small k, large p/k, and moderate sample sizes.

Lastly, the bias estimates for conditions in which the q variables evidenced focused but inconsistent correlations with the p variables are presented in Table 6. The performance of the SW method with this type of anomalous variable was notably different from the performance observed with the other anomalies examined. For this particular anomaly, the SW method evidenced more pronounced positive bias in the resulting structure coefficients than that obtained under the other two anomalous conditions. However, the bias for the SW method was still substantially less than the bias witnessed for the TA method. For the SW method, bias was reduced with larger matrices (both for more variables per factor and for larger numbers of factors). For the TA method, the bias pattern was consistent with the pattern observed for the other types of anomalies (with the largest bias observed for conditions with small k, large p/k, and moderate sample sizes).

RMSE

In addition to estimates of statistical bias, the root mean squared errors (RMSEs) of the factor structure coefficients of the SW and TA methods are provided in Tables 4–6. These statistics reflect sampling variability in terms of squared deviations from the population parameter. The average value of the RMSE of sample structure coefficients for both the p variables and q variables are presented, as well as an overall estimate. If a statistic is unbiased, the RMSE is the same as the standard error. Because these statistics reflect sampling error, it may be expected that the RMSEs would become smaller with larger sample sizes (e.g., for conditions with a large N per variable). When estimators are biased, however, the RMSE may not decrease with larger sample sizes.

The RMSE of the p variables was consistent in magnitude for both the SW and TA methods across virtually all conditions. These estimates were seen to decrease dramatically with larger sample sizes and larger matrices. For example, with $k = 2$, $p/k = 5$, $q = 2$, the RMSE of the p variables decreased from approximately .11 to .04, across the three sample sizes examined (Table 4). This pattern was also consistent across the three anomalous conditions.

For the q variables, the magnitude of the RMSE associated with the factor structure coefficients for the TA method generally exceeded those estimates for the SW method (as expected, because of the greater bias in the TA estimates). Although sampling error was substantially reduced for both methods with an increase in sample size and number of factors, this reduction was somewhat more pronounced for the SW method. For example, with $k = 2$, $p/k = 3$, and $q = 2$, RMSE ranged from .20 to .06 for the SW method and from .30 to .20 for the TA method across the three sample sizes examined (Table 4).

Discussion

Although this empirical comparison of a stepwise selection algorithm to a traditional exploratory factor analysis approach revealed many similarities in the resulting factor solutions, two major areas of difference were evident: selection accuracy and statistical bias in factor struc-

TABLE 5.
Statistical bias, RMSE and pattern accuracy for stepwise and traditional exploratory factor analyses [anomaly = low level dispersed correlation].

				Bias						RMSE						Pattern accuracy			
				p Vars		q Vars		Overall		p Vars		q Vars		Overall		Per item		All items	
k	p/k	q	n/Var	SW	TA	SW	TA	SW	TA	SW	TA	SW	TA	SW	TA	SW	TA	SW	TA
2	3	1	5	−2	−1	−1	1	−1	0	18	17	19	21	19	19	93	92	70	68
2	3	1	10	−2	−2	1	11	−2	0	11	11	17	26	12	13	98	98	91	89
2	3	1	50	−2	−2	−2	3	−2	1	6	7	6	13	6	10	100	99	97	94
2	3	2	5	−2	−1	−1	2	−1	1	17	17	20	22	19	20	94	93	73	70
2	3	2	10	−2	−2	1	11	−1	1	11	11	16	24	12	14	99	98	93	90
2	3	2	50	−2	−2	−1	5	−2	2	6	6	6	16	6	12	100	99	98	94
2	5	1	5	0	0	0	2	0	1	12	12	13	15	12	13	99	98	87	85
2	5	1	10	−1	−1	0	15	−1	1	8	9	12	27	8	10	100	99	96	94
2	5	1	50	−1	−1	−1	3	−1	1	4	4	4	10	4	7	100	99	97	95
2	5	2	5	0	0	0	3	0	2	11	12	13	15	12	14	99	98	90	87
2	5	2	10	−1	−1	0	13	0	2	8	8	11	23	8	11	100	99	97	94
2	5	2	50	−1	−1	−1	5	−1	2	4	4	4	13	4	9	100	99	97	95
2	10	1	5	0	0	0	2	0	1	8	8	8	10	8	9	100	100	95	93
2	10	1	10	0	0	0	19	0	1	6	6	8	35	6	7	100	100	96	94
2	10	1	50	0	0	0	2	0	1	3	3	3	6	3	4	100	100	97	96
2	10	2	5	0	0	0	2	0	1	8	8	8	11	8	10	100	99	96	94
2	10	2	10	0	0	0	16	0	1	5	6	7	26	6	8	100	100	97	95
2	10	2	50	0	0	0	3	0	2	2	3	3	8	3	6	100	100	97	95
5	3	1	5	−1	0	1	3	0	0	11	11	17	20	12	12	98	98	80	79
5	3	1	10	−1	−1	1	5	−1	0	7	8	11	19	8	9	100	99	97	95
5	3	1	50	−1	−1	0	5	−1	−1	4	4	5	23	4	6	100	99	96	94
5	3	5	5	−1	−1	0	3	0	0	11	11	15	18	12	13	98	97	82	79
5	3	5	10	−1	−1	0	4	0	1	7	7	10	15	8	9	100	99	98	94
5	3	5	50	−1	−1	0	5	−1	0	4	4	4	23	4	7	100	99	98	94
5	5	1	5	0	0	0	5	0	0	8	8	11	20	8	9	100	99	96	94
5	5	1	10	0	0	0	6	0	0	6	6	8	25	6	7	100	99	96	95
5	5	1	50	0	0	0	5	0	0	3	3	3	24	3	4	100	99	97	96
5	5	5	5	0	0	0	5	0	1	7	8	10	16	8	9	100	99	97	95
5	5	5	10	0	0	0	5	0	1	5	6	7	16	6	7	100	99	98	95
5	5	5	50	0	0	0	5	0	0	2	3	3	21	3	5	100	99	98	95
5	10	1	5	0	0	0	6	0	0	5	6	7	25	5	6	100	100	96	94
5	10	1	10	0	0	0	6	0	0	4	4	5	25	4	4	100	100	96	96
5	10	1	50	0	0	0	6	0	0	2	2	2	25	2	2	100	100	97	96
5	10	5	5	0	0	0	5	0	1	5	5	7	17	5	6	100	100	97	94
5	10	5	10	0	0	0	5	0	0	4	4	5	25	4	5	100	100	97	95
5	10	5	50	0	0	0	6	0	0	2	2	2	20	2	3	100	100	98	96

Note. Values have been multiplied by 100 and rounded. Estimates are based on 10,000 replications.

TABLE 6.
Statistical bias, RMSE and pattern accuracy for stepwise and traditional exploratory factor analyses [anomaly = focused inconsistent correlation].

				Bias						RMSE						Pattern accuracy			
				p Vars		q Vars		Overall		p Vars		q Vars		Overall		Per item		All items	
k	p/k	q	n/Var	SW	TA	SW	TA	SW	TA	SW	TA	SW	TA	SW	TA	SW	TA	SW	TA
2	3	1	5	−2	−1	0	2	−1	1	18	18	21	23	20	21	92	91	67	65
2	3	1	10	−2	−2	7	15	−1	1	12	12	20	32	13	15	98	97	88	86
2	3	1	50	−2	−2	0	5	−1	2	6	7	7	15	7	11	99	99	94	94
2	3	2	5	−2	−2	1	4	0	2	18	18	22	25	20	22	92	91	68	65
2	3	2	10	−2	−2	6	14	0	2	11	12	19	30	13	17	98	97	90	87
2	3	2	50	−2	−2	1	7	0	4	5	7	8	16	7	13	99	98	94	94
2	5	1	5	0	0	0	2	0	1	12	12	13	16	13	14	98	98	86	85
2	5	1	10	−1	−1	4	22	0	1	8	9	13	36	9	11	99	99	95	94
2	5	1	50	−1	−1	0	4	0	2	4	4	4	11	4	8	100	99	95	95
2	5	2	5	0	0	1	3	0	2	12	12	13	17	13	15	99	98	88	86
2	5	2	10	−1	0	4	17	0	2	8	8	13	31	9	12	100	99	95	95
2	5	2	50	−1	−1	0	4	0	2	4	4	5	11	4	8	99	99	95	95
2	10	1	5	0	0	0	3	0	1	8	8	8	11	8	10	100	99	94	93
2	10	1	10	0	0	2	30	0	1	6	6	8	43	6	8	100	100	95	95
2	10	1	50	0	0	0	3	0	1	3	3	3	6	3	5	100	100	96	96
2	10	2	5	0	0	0	3	0	2	8	8	8	12	8	10	100	99	94	93
2	10	2	10	0	0	2	27	0	2	5	6	8	42	6	9	100	100	95	95
2	10	2	50	0	0	0	3	0	1	2	3	3	6	3	5	100	100	95	95
5	3	1	5	−1	−1	3	4	0	0	12	12	19	22	12	12	97	97	76	75
5	3	1	10	−1	−1	3	6	0	0	7	8	13	21	8	9	100	99	95	95
5	3	1	50	−1	−1	3	11	−1	0	4	4	7	25	4	6	100	99	94	94
5	3	5	5	−1	−1	3	4	0	0	12	12	17	21	13	14	96	95	71	67
5	3	5	10	−1	−1	3	6	0	1	7	8	12	18	8	10	100	99	95	94
5	3	5	50	−1	−1	2	11	0	1	3	5	6	24	4	7	99	98	93	93
5	5	1	5	0	0	1	7	0	0	8	8	12	23	8	9	100	99	94	94
5	5	1	10	0	0	2	11	0	0	6	6	9	27	6	7	100	99	95	95
5	5	1	50	0	0	1	12	0	0	3	3	4	26	3	4	100	99	95	95
5	5	5	5	0	0	1	4	0	1	8	8	11	19	8	10	100	99	94	94
5	5	5	10	0	0	1	7	0	1	5	6	8	18	6	8	100	99	95	95
5	5	5	50	0	0	1	12	0	1	2	3	4	20	3	4	100	99	95	95
5	10	1	5	0	0	1	12	0	0	5	6	7	28	5	6	100	100	95	94
5	10	1	10	0	0	1	12	0	0	4	4	5	27	4	4	100	100	95	95
5	10	1	50	0	0	1	12	0	0	2	2	3	27	2	2	100	100	96	96
5	10	5	5	0	0	1	6	0	0	5	5	7	18	5	6	100	100	94	94
5	10	5	10	0	0	1	12	0	0	4	4	5	27	4	4	100	100	96	96
5	10	5	50	0	0	1	12	0	0	2	2	3	27	2	2	100	100	96	96

Note. Values have been multiplied by 100 and rounded. Estimates are based on 10,000 replications.

ture coefficients. In selection accuracy, both methods were comparable in sensitivity across the conditions examined (i.e., both methods identified the p variables that were well explained by the k common factors). However, the stepwise method showed poor specificity (the ability to exclude the q variables that were not well explained by the factors) for all but a limited number of conditions. Specifically, the stepwise method showed virtually no specificity with anomalous variables that evidenced either no correlation with the other variables or a low-level diffuse correlation. Further, this lack of specificity was evident across all sample sizes examined. Only with anomalous variables that produced focused but inconsistent correlations did the stepwise method exclude the q variables from the final solution. For this type of anomalous variable, the power of the test of fit was a critical factor in the resulting specificity (i.e., specificity increased with larger samples).

Such a result was anticipated because the fit statistic used to drive the stepwise approach considers a rather narrow definition of an "anomalous variable." Many variables that would appear to be a poor choice for inclusion in a factor analysis (e.g., variables that do not correlate with the other variables in the matrix or variables that correlate at a low level with all other variables) do not impact the degree of correspondence between the observed and implied covariance matrices. As evidenced in these simulations, such variables are not adequately screened out by the stepwise algorithm.

Despite the poor specificity of the stepwise method, the sample factor solutions obtained by this method evidenced less bias in the structure coefficients than those obtained by the traditional method. The increased bias seen in the traditional method, however, was localized in the q variables. Such an increase in bias was also anticipated because of the particular inclusion criterion that was employed in the traditional approach as operationalized in this study. Variables were only selected for inclusion in a factor solution if the absolute value of the sample factor loading exceeded .30. Thus, with the traditional approach, the samples with smaller loadings (both positive and negative in direction) were not included in the factor solution and did not contribute to the bias calculations. Because the stepwise method was not constrained to exclude the q variables if the loadings were small in magnitude, the biases in the loadings resulting from the use of this method were typically close to zero.

The interpretation of these results needs to be considered in the context of the limitations of this research. Specifically, all of the simulations were conducted with orthogonal factor solutions in which the number of factors was predetermined. The decision to examine only orthogonal factor solutions was made in order to provide an initial evaluation of the performance of the stepwise method under rather simple conditions. If the stepwise method had evidenced superior performance under the conditions currently examined, then exploration of oblique factor solutions would appear warranted. The failure of the stepwise method to perform adequately under such parsimonious conditions essentially circumvents the need to examine more complex, sophisticated models or conditions.

In the conduct of actual exploratory factor analyses, a researcher must decide on the number of factors to extract and on the form of rotation that is most appropriate. Further, the population matrices for the p variables were constructed to represent perfectly simple structure, with each variable being a function of only a single factor and with identical loadings present for each variable. Routinely, social science data are characterized by less clear structures and the performance of any exploratory factor analysis technique is likely to be less optimistic than results obtained in this study. Additionally in these simulations, only three types of anomalous variables were examined, and for no conditions were the types of anomalous variables mixed. Since the performance of the stepwise algorithm depended on the type of anomalous variable, one may question how the algorithm would perform with other types of anomalous variables, such as those that have correlations with other variables that can be partially, but not fully explained by the factor structure.

In addition to broadening the range of data contexts examined, future research should consider the details in the application of the variable selection methods. The initial factor solution that is required for the stepwise algorithm was always the "correct" one (i.e., the stepwise method began with a factor solution that included all of the p variables and none of the q variables). The performance of the stepwise algorithm may deteriorate under less optimal starting conditions. Additionally the steps in the stepwise algorithm were based on an overall chi-square test of fit. More traditional stepwise algorithms are based on incremental changes, and the performance of the stepwise algorithm may be altered if steps were responsive to incremental changes in fit. It would also be interesting to see how performance would respond to shifting focus from a chi-square test of statistical significance to a descriptive fit index, such as the Comparative Fit Index (CFI; Bentler, 1990) or the Root Mean Square Error of Approximation (RMSEA; Browne & Cudeck, 1993). Again overall levels of fit or incremental changes in fit could be considered.

Finally, in the traditional factor analysis approach, only a single rule of thumb was employed. We recognize and appreciate the difficulty of developing an objective set of rules to capture the multifaceted heuristics that researchers use in applied exploratory factor analysis. However, our decision to apply this single rule of thumb was deemed appropriate because we have found that the .30 saliency rule is the most commonly used method in a variety of fields in the social sciences. Improved results for the traditional approach may be found for rules that incorporate the influence of sample size through explicit consideration of anticipated sampling error of the factor loadings (see, for example, Cliff & Hamburger, 1967; Cudeck & O'Dell, 1994). Because these rules require substantially larger loadings than .30 for small sample analyses, they may be expected to provide improved specificity in such conditions. Similarly, the use of more elegant rules (such as a "simple structure" rule, requiring a substantial loading on only a single factor) may yield higher specificity than was obtained with the simple .30 rule applied in this research.

In summary, although the stepwise selection approach examined here performed better on many criteria than stepwise approaches used in regression and discriminant function analyses, our results do not support its use in exploratory factor analysis. An important distinction is that stepwise approaches used in the latter applications seek to maximize a relationship with a minimum number of variables, thus capitalizing on chance relationships that occur in samples. In contrast, the stepwise selection approach for exploratory factor analysis seeks to include a maximum number of variables without substantially deteriorating the fit index. Despite the better than anticipated performance of the stepwise procedure in terms of criteria such as statistical bias and RMSE in the factor structure coefficients, the results suggest that the stepwise procedure is generally ineffective at excluding anomalous variables from the factor model. The stepwise algorithm only evidenced reasonable exclusions for a limited type of anomaly (variables with a focused correlation inconsistent with the k-factor model) and only for large sample sizes. The poor selection accuracy of the stepwise approach suggests that it should be avoided. In the majority of conditions examined, better factor solutions were obtained from the simpler traditional approach.

References

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238–246.

Browne, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K.A. Bollen & J.S. Long (Eds.), Testing structural equation models (pp. 136–162). Newbury Park, CA: Sage.

Cattell, R.B. (1978). *The scientific use of factor analysis in behavioral and life sciences*. New York: Plenum.

Cliff, N., & Hamburger, C.D. (1967). The study of sampling errors in factor analysis by means of artificial experiments. *Psychological Bulletin, 68*, 430–445.

Comrey, A.L. (1973). *A first courses in factor analysis*. New York: Academic Press.

Cudeck, R., & O'Dell, L. (1994). Applications of standard error estimates in unrestricted factor analysis: Significance tests for factor loadings and correlations. *Psychological Bulletin, 115*, 475–487.

Everitt, B.S. (1975). Multivariate analysis: The need for data, and other problems. *British Journal of Psychiatry*, *126*, 237–240.

Fabrigar, L.R., Wegener, D.T., MacCallum, R.C., & Strahan, E.J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272–299.

Finch, J.F., & West, S.G. (1997). The investigation of personality structure: Statistical models. *Journal of Research in Personality*, *31*, 439–485.

Gorsuch, R.L. (1973).Using Bartlett's significance test to determine the number of factors to extract. *Educational and Psychological Measurement*, *33*, 361–364.

Gorsuch, R.L. (1988). Exploratory factor analysis. In J.R. Nesselroade & R.B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 231–258.) New York: Plenum.

Huberty, C.J. (1989). Problems with stepwise methods—better alternatives. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 1, pp. 43–70). Greenwich, CT: JAI Press.

Kano, Y., & Harada, A. (2000). Stepwise variable selection in factor analysis. *Psychometrika*, *65*, 7–22.

Knapp, T.R. (1978). Canonical correlation analysis: A general parametric significance testing system. *Psychological Bulletin*, *85*, 410–416.

Little, T.D., Lindenberger, U., & Nesselroade, J.R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When "good" indicators are bad and "bad" indicators are good. *Psychological Methods*, *4*, 192–211.

MacCallum, R.C., Roznowski, M., & Necowitz, L.B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*, 490–504.

MacCallum, R.C., Widaman, K.F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*, 84–89.

Nunnally, J.C. (1994). *Psychometric Theory* (3rd ed.). New York: McGraw-Hill.

Robey, R.R., & Barcikowski, R.S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology*, *45*, 283–288.

Snyder, P. (1991). Three reasons why stepwise regression methods should not be used by researchers. In B. Thompson (Ed.), *Advances in educational research: Substantive findings, methodological developments* (Vol. 1, pp. 99–106). Greenwich, CT: JAI Press.

Tabachnick, B.G., & Fidell, L.S. (1983). *Using multivariate statistics* (2nd ed.). New York: Harper & Row.

Thompson, B. (1988a, April). *Canonical correlation analysis: An explanation with comments on correct practice*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED 295 957).

Thompson, B. (1988b, November). *Common methodology mistakes in dissertations: Improving dissertation quality*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville, KY. (ERIC Document Reproduction Service No. ED 301 595).

Thompson, B. (1989). Why won't stepwise methods die? *Measurement and Evaluation in Counseling and Development*, *21*, 146–148.

Thompson, B. (1991, January). *Stepwise methods lead to bad interpretations: Better alternatives*. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX. (ERIC Document Reproduction Service No. ED 327 573).

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement*, *55*, 525–534.

Thurstone, L.L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.

Velicer, W.F., & Fava, J.L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods*, *3*, 231–251.

Manuscript received 5 JUN 2001
Final version received 25 AUG 2003